DOCUMENT RESUME

ED 353 358 UD 029 033

AUTHOR Sweet, David

TITLE Reconsidering Current Federal Policy for Evaluating ESEA Chapter 1 Grants to Local Education Agencies.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO TAC-B-301
PUB DATE Apr 91
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; *Compensatory Education; Course Content; *Economically Disadvantaged; Educational Assessment; *Educational Policy; Educational Quality; Elementary Secondary Education; Federal Legislation; *Federal Programs; Grants; *Measurement Techniques; Outcomes of Education; Policy Formation; *Program Evaluation; School Districts
IDENTIFIERS Education Consolidation Improvement Act Chapter 1; Hawkins Stafford Act 1988

ABSTRACT

This paper presents arguments for changing a portion of current federal policy for evaluating Chapter 1 grants to Local Education Agencies (LEAs). Chapter 1, part of the federal Elementary Secondary Education Act (ESEA), provides funds to schools and districts to improve education for economically disadvantaged students. The paper advocates that multiple measures and some local selection of appropriate measures of program quality be used in addition to appropriate achievement outcome measures to determine which programs need school improvement. In addition, alternatives to norm-referenced standardized tests should be permitted as achievement outcome measures. In making the argument for these changes, the paper reviews norm-referenced tests, types of achievement performance missing from current tests, subject matter content missing from current tests, what the new performance assessments are, details of the current policy for program evaluation, and the LEAs' response to these provisions. Later sections argue that this is a good time to rethink evaluation requirements, that the current policy is less effective, and that performance assessments should be available for use in Chapter 1 programs. A final section discusses how to balance a program-specific array of instruments for evaluation. Included is a 12-item bibliography. (JB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED353358

TAC B-301

Reconsidering current Federal policy for evaluating ESEA Chapter 1 grants to Local Education Agencies

by
David Sweet
U. S. Department of Education
O.E.R.I.
Office of Research

presented to
the
American Educational Research Association
April 1991

U.S. Department of Education Official Disclaimer: This paper is intended to promote the exchange of ideas among researchers and policy makers. The views are those of the author, and no official support by the U.S. Department of Education is intended or should be inferred.

UD029033

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

BEST COPY AVAILABLE

Table of Contents

A.	Introduction	1
B.	Norm-referenced standardized tests.	2
C.	What types of achievement performance are current tests missing?	3
D.	What subject matter content is missing in current tests?	5
E.	What are the new performance assessments?	7
F.	What precisely is the current policy for Chapter 1 program evaluation?	10
G.	What has been the LEA response to these provisions?	11
H.	A time for rethinking Chapter 1 requirements.	12
I.	Criticism of the current evaluation policy.	13
J.	Why should some performance assessments be available for use in Chapter 1 programs?	14
K.	In search of a better array, balance and use of item formats	16

Reconsidering current Federal policy for evaluating ESEA Chapter 1 grants to LEAs

A. Introduction

Mothers and fathers across America send their children to school each morning with the hope that their children will learn about this country and its place in the world at large; gain an appreciation of the letters, arts and sciences; develop interests in work and further learning; acquire skills and habits of behavior and communication that are appropriate for a variety of settings; and grab hold of the keys to unlock doors that may stand between them and the American dream. Not all students start this journey that is education with the same preparation or support. It is because of these differences among students that we have special programs. The largest single Federal program in education, known as Chapter 1 of the Elementary and Secondary Education Act (ESEA), is intended to supply additional support and services to those who may need services more than others,-- namely low achieving students living in low income areas.

I am here today to advocate a change in a small part of the current Federal policy of Chapter 1. It is a small part of the program but one that has potential for improving several aspects of Chapter 1 programs and the quality of education Chapter 1 students need and deserve.

In particular, I am advocating a change in the Federal policy for evaluating Chapter 1 grants to Local Educational Agencies (LEAs). The current policy requires LEAs to initiate a school improvement program in any school where the Chapter 1 students fail to demonstrate improvement relative to others on a norm-referenced standardized test.

While at first glance the requirement to use these standardized tests may appear to be a minor, technical, perhaps trivial, issue, it is not. Research shows that such a policy statement requiring particular tests is anything but inconsequential (Sheppard, 1989; Madaus, 1988). The same conclusion has been drawn from a systems analytic perspective (Fredericksen & Collins, 1989).

What I am advocating is the following:

- o No single measure should be made the exclusive Federal criterion for evaluating and determining which Chapter 1 programs need school improvement. Multiple measures are needed for complex decisions; there should be some flexibility for local selection of appropriate measures of program quality that could be considered in addition to appropriate achievement outcome measures.

- o Alternatives to norm-referenced standardized tests, especially those with strong performance assessment components, should be permitted and encouraged as achievement outcome measures in evaluating Chapter 1 programs.

Development of an argument in support of this position will require some excursions into several areas, including: [list = headings of subsequent sections of this paper]

B. Norm-referenced standardized tests.

Norm-referenced standardized multiple choice tests can be constructed to measure factual knowledge, skills, understanding, reasoning, and even "higher-order skills" in reading, writing, mathematics, science, history, literature, geography, foreign languages, just about any subject that might come up in an elementary or secondary school curriculum. The readability of the items can be set for any number of age, grade or developmental levels. They are available in a range of print sizes, styles and colors. They come with answer sheets, documentation about what topics are covered, and materials on how to interpret results and share the results with parents. Although they are often administered by the classroom teacher, scoring and analysis is a mechanical/algorithmic process carried out off the school grounds, usually by the test publisher. The process is as aseptic and credible as sending a blood sample to the laboratory. Although test scoring takes more time than the turnaround by medical laboratories, test administration takes less time than a typical visit to the doctor's office. And, they are cheap, costing usually between $2.00 and $4.50 per student depending on the test, services and number of students tested. These tests are all over the place; they're popular, as common as Snickers bars. And for the record, I love Snickers bars.

So, what's the problem? The problem is that despite all the good things that norm-referenced standardized tests can do and despite the good intentions of test publishers and sophisticated methods used in test

construction and analysis, there are very important parts of education that are being missed and others that are being actually harmed by these tests and the role they have taken on in today's testing environment. The harm is not so much in these tests themselves; the harm is in using them excessively and in situations where something else is needed.

Too often the commendable features of multiple choice tests -- objectivity in scoring, low cost, and seeming flexibility to meet needs in all situations -- have been oversold or overbought. Too often and too long, multiple choice tests have been the answer, the quick, convenient fix to too many of our problems.

EXAMPLE: Find that teachers don't mark papers in the same way? Then give a multiple choice test which has "correct" answers. The multiple choice test strategy, however, solves only part of the problem and creates one of its own: The problem left unsolved is that teachers still don't receive training to consistent standards of marking papers. The added problem is that the rating and standard setting process is separated from teachers and the instructional process.

C. What types of achievement performance are current tests missing?

It is an old saw that you have to know what you want before you ask for it. Consider some the things we would like to know about student achievement, what students know and can do. A partial list might include the following:

- Depth of knowledge,
- Breadth of knowledge,
- Ability to selectively use knowledge and skills,
- Ability to structure problems and setup solution strategies,
- Ability to recognize quality work in self and others,
- Improvement,

- o Speed,
- o Effort and sustained effort,
- o Individual and group productivity,
- o Attitude and interest, and
- o Various cognitive structures and processes used in problem solving.

All of these items can be measured with appropriate measurement instruments and procedures. All but the last of these items are outcome measures, aspects or dimensions of outcome performance. I included the last item on the list as a reminder of an additional realm of cognitive measures that can be used by skilled teachers in managing instruction for groups and individual students. Although they and attitudes and interests can be achievement outcome measures, they are at least as often thought of as an interim rather than a final outcome measure.

Ask yourself why we do not measure most of these phenomena. I think you will find that the answers involve technical and resource issues more than educational ones. They are also likely to include some pat responses, which really need some rethinking.

EXAMPLES: We don't have reliable measures for that. We can't measure both speed and power on the same test. What do you mean by "effort" and how do you measure that in a test? "Improvement" -- That's a pretest and posttest isn't it? Why should we be interested in measuring speed? I can't measure productivity -- everybody gets the same assignment.

Some things we currently do not measure, not because they are educationally unimportant, but because of technical considerations in measurement. Measurement driven testing. Sounds harmless doesn't it? One problem with allowing measurement issues to drive testing practice is that the resultant testing misses a lot. Teachers have less information to work with; students get less useful feedback; and policymakers receive less valid information than the data labels on reports suggest. Another problem (discussed in the next section) is that these tests limit the subject matter assessed to decontextualized, decomposed fragments; this, in turn, has what

only can be called pernicious consequences for what is taught and learned. An equally serious problem (discussed in a later section) is that the convenient, cheap measurement driven tests are used as a policy action and leave the impression that the underlying problems have been addressed.

D. What subject matter content is missing in current tests?

Besides limiting the types of achievement measures we obtain, current testing practices limit the subject matter domains that are actually assessed. These tests generally cover only the knowledge and skills that can be measured in multiple choice questions. By the very test construction requirements used in developing norm referenced standardized tests, the items are problems that have little or no context (decontextualized) and measure only fragmented pieces of knowledge (decomposed).

Once promoted as the basic building blocks for fuller understanding, isolated skills and bits of knowledge often remain precisely what they are, never integrated or applied, until they are finally forgotten.

Bransford and Vye (1989) have shown that students bring their own individual concepts and intuition to school at very young ages. Despite the fact that they learn facts, symbols and algorithms in class, students tend to revert to primitive, *unaltered* concepts and intuition in all but the *structured* class/test settings. Resnick (1989) has shown that thinking skills are intimately involved in successful learning of even elementary levels of reading, mathematics and other subjects. She and a number of curriculum development experts and Neo-Piagetian cognitive psychologists have stressed the need for children to develop understanding through their own constructive activities. Without this individual constructive activity of confronting new observations and ideas with their preexisting ideas, there is no development. There is no accommodating the new ideas into an individual's thinking to understand the circumstances under which the new ideas do or do not apply. And there is no adaptation of replacing or modifying the original ideas. The analogy is with phagocytosis, the process by which a cell breaks down and assimilates what it has ingested: If the new idea is not broken down and placed in context, it remains separate and does not nourish thinking, and it will eventually be expelled. In the classroom and testing context, this takes on a variety of forms. Numerous researchers have reported how students regurgitate facts, repeat interpretations they believe the teacher wants, plug in supplied numbers to formulas presented in the current chapter, etc. Song writers are even more succinct: "When I think back to all the crap I learned in high school, it's a

wonder I can think at all."(Paul Simon)

UNSUPPORTED CLAIMS: Many of the current tests that are composed almost entirely of multiple choice items claim to measure problem solving ability and higher order thinking skills as well as a range of basic and advanced skills. This claim makes some serious assumptions about the cognitive activity students engage in while they are taking tests. Although students could construct their own individual solutions to the problem inherent in the stem of the question and then compare their own solution to the choices provided by the test developer, there is little evidence that this strategy is either used often or conducive to maximizing performance, except in simple computational types of problems.

The independent construction strategy may in fact lead to "fighting the test" (mentally arguing with how the wording of the correct solution that is provided) in all but the simplest questions and getting fewer items completed within a given time period. A more efficient strategy may in fact involve problem solving of a very different sort from that described in the test materials,-- i.e. looking for a reasonably correct solution statement (and possibly checking that none of the other choices is a more complete or better statement) and eliminating incorrect or otherwise inferior choices.

How difficult an individual item is and how much it counts for in an IRT scale score for a student does not depend solely on the difficulty or complexity of the cognitive problem posed by the item itself. Instead, item difficulty depends on a host of other particulars, including: the clarity/obfuscation with which the problem is presented; the linguistic demands of the text, both the stem and the choices provided; the student's familiarity (exposure or opportunity to learn) the subject matter and other references that make up the item; the pattern of choices,-- their distinctiveness regardless of whether the distinctions are rooted in the problem itself, closely related subject matter knowledge, general aptitude or other factors. The functioning of an item in the test also depends on: its location or position in the test; the nature of the other items in the test, and the conditions (time, instructions, consequences for student) under which the test is administered.

Multiple choice items have fixed, inflexible levels of difficulty and, except for the rare situation of adaptive testing, test items are given to students without regard to mismatches between item difficulty levels and student ability levels. Thus, unlike life outside testing situations where problems

and opportunities can be addressed more or less well -- the problems in multiple choice tests have only one "correct" solution.

EXAMPLES: There is a big difference in feedback to students. For multiple choice tests, it is: "You got this, this and this right, and this and this wrong." For tests with tasks that students can perform more or less well, it is: "Here is the answer you gave, and here are three ways you can improve your answer." One form is not always better, even if it is cheaper and easier, and that is the nub of the problem with the multiple choice monopoly.

From a measurement perspective, this single correct solution paradigm means that the whole array of potential solutions (e.g.-- superior, equivalent and less adequate solutions) are mapped into two points -- "correct" and "not correct". Beyond measurement problems, the paradigm has more troubling implications for students intellectual standards, their curiosity and interests, their self concepts, and their habits of thinking and working. These are the most pernicious consequences of the multiple choice format. They have been discussed elsewhere (Resnick, 1989; Schwartz & Viator, 1990), which I recommend to you.

Moreover, the mapping is far from perfect. The fact that a student can generate his or her own accurate solution to a problem (i.e., the student can *accurately describe* the author's point of view), is no guarantee that he or she will be successful in choosing the correct provided response (i.e., the student may find the test developer's description of the author's point of view to be deficient and opt for the choice "none of the above".) [The test publisher's only check on this quirk is by looking at how students of varying estimated proficiency levels and demographic characteristics respond to the provided choices.] Similarly, the fact that a student cannot generate his or her own solution to a problem is no guarantee that he will be unsuccessful in choosing the correct provided response. He or she might, though unable to recall without a cue, be able to recognize the correct solution as an accurate one. Or, he or she could choose the correct solution through inaccurate/faulty reasoning or guessing.

E. What are the new performance assessments?

Multiple choice tests are not the devil incarnate. It is their dominance, their predominance, their tendency to drive out other forms of tests -- that is the problem. And, it is in the recognition of the two way linkages between curriculum and testing that education reformers have found a point of leverage to get us out of our rut. The strategy is to change to a better, more thoughtful curriculum by changing to tests that require more thinking. The hope is that the new assessments will serve as an explicit tool for setting standards and will, along with other ingredients of curriculum development and school reform, make both students and teachers active agents of the change process.

The term "performance assessment" has long been used to describe a category of assessments that require students actually to perform, demonstrate, construct and/or develop a product, solution, etc. under defined conditions and standards, compared with most current tests that require students to select from among descriptions of solutions that have been provided by the test developers. They represent an interest in valuing and measuring how students can use knowledge in a real world context. "Portfolios" are a special form of the new assessments, consisting of collections of a student's work showing a variety of exemplary final products as well as some works possibly at several stages of development.

Numerous other assessment forms are being researched and tried out. For example, performance assessments also include: exhibits such as those in science fairs; expanded projects that often include the work of several students; and a collections of procedures called expanded or enhanced multiple choice tasks that have various ways of ensuring that a student generates his or her own solution to a problem before considering multiple choice options.

Rooted in a master's assessment of apprentice work, performance assessments have an established role in licensing examinations for several professions, both in and outside this country. They also have a tradition in many of the older colleges and universities. Similar procedures have been used in civil service examinations in other countries and are now being used increasingly at the secondary level in Europe. The new assessments are "new" in the sense that they are being used for the first time on a wide scale to measure achievement performance in academic subjects in elementary and secondary schools in the United States.

These new assessment of student performance are seen as playing a key role in school reform in at least two respects. First, improved student performance has been singled out as one of school reform's goals; and assessment, broadly conceived, is recognized as a collection of methods for

measuring student performance. School reform is not only striving to bring about new levels of performance, but is also endeavoring to advance and expand our very understanding of the behaviors, knowledge, and skills that need to be valued (and therefore assessed) in the education system. The expansion to new types of performance requires new performance measures, new assessments.

Second, assessment has been identified as a tool or instrument to bring about (and later sustain) many of the changes central to the school reform movement. Motivated by the potential leverage that a transformed assessment system could have on curriculum and instruction and a recognition that the existing traditional assessments would continue to impede efforts to reform curriculum, school reform proponents see the need for assessment reform. Where students, teachers and schools are held accountable on the basis of student performance in subjects XYZ, those subjects will valued and the curriculum and instruction will be heavily tilted toward subjects XYZ. In a like manner it is argued that, where assessments include problem-solving in a real world context without extensive structuring of solution strategies and without provided multiple choice answers, student problem-solving will be valued and appear prominently in the curriculum and instruction. Where sustained student effort is not part of the assessment and plays no role in the accountability, it is valued less and is emphasized less in the curriculum and instruction.

Assessment reform is also seen as a potentially powerful tool for improving instruction and raising standards of performance; the new assessments can provide more integrated, more concrete information about what real problems look like and how a student's preliminary solution can be improved. This information is useful for both students and teachers and compares favorably with that provided by more traditional assessments.

These new forms of assessments, designed to measure student performance on a broader, balanced and more integrated collection of knowledge, skills and behaviors, are being developed and implemented across the country. A recent survey (Pelavin, 1990) reports that 33 States are already trying out some form of performance assessment in their own assessment program and that additional States are engaged in development work on performance assessments. California, Connecticut, New York, Massachusetts, Maryland and Vermont have been particularly active in this area. Information about which school districts and individual programs are currently developing and implementing these new assessments is fragmented and incomplete.

F. What precisely is the current policy for Chapter 1 program evaluation?

The reason these concerns with current tests and testing policies need to be considered in the context of Chapter 1 evaluations is because Federal law, regulations and policy say that school districts must use norm-referenced standardized tests to evaluate local school Chapter 1 programs. Moreover, under current policy, schools are identified for intervention if these test results do not meet Federal standards, regardless of what other information (achievement or otherwise) they have to the contrary.

The policy is stated most explicitly in the Federal Register Rules and Regulations issued May 19, 1989 (Vol. 54 No. 96) Subpart H --What Are the National Evaluation Standards? The pertinent sections are 200.82 (procedures) and 200.83 (alternative procedures or exceptions).

> **Section 200.82 What procedures does an LEA use in evaluating student achievement?** Unless it is using approved alternative procedures under Section 200.83, an LEA shall use the following procedures to evaluate student achievement in each Chapter 1 project funded under this part that provides instructional services in reading, language arts, or mathematics in grades 2 through 12 during the regular school year.
> (a) The LEA shall administer a pretest and a posttest separated by approximately 12 months.
> (b) The LEA may use a test with or without national norms as follows:
>
> > (1) If the LEA uses a test with national norms, the LEA shall administer the test within the appropriate range of the test publisher's norming dates.
> > (2) If the LEA uses a test without national norms, the LEA shall adhere to technical requirements for equating this test with a nationally normed test as specified by the Title I Evaluation and Reporting System or other valid methods accepted by the Secretary.
>
> **Section 200.83 What alternative procedures may an LEA use?** (a) An LEA may use alternative procedures to those in Section 200.82 for evaluating student achievement if, before using the alternative procedures, the LEA obtains the approval of, first, the SEA, and then the Secretary.
> (b) In order for the SEA and the Secretary to approve alternative procedures, the LEA shall demonstrate that the procedure--
>
> > (1) Yield a valid and reliable measure of--

(i) The Chapter 1 children's performance in reading, language arts, or mathematics; and
(ii) The children's expected performance; and
(2) Produce results that can be expressed in the common reporting scale established by the Secretary for SEA reporting. (Chapter 1 Policy Manual, 1990)

G. What has been the LEA response to these provisions?

As of December 1990 the Department's Office of Compensatory Education had conducted a total of 43 jurisdiction reviews of the administration of Chapter 1 programs by State Education Agencies in 40 of the 50 States with the remaining 3 reviews being of programs administered by the District of Columbia, the Bureau of Indian Affairs and Puerto Rico.

Data from these reviews indicate that there has been a high level of State and local compliance with the evaluation provisions. All 43 jurisdictions had adopted standards expressed on a scale of norm curve equivalent (NCE) gains between pretests and posttests.

According to the Department report:

> Four States had initially adopted more complex standards, either by allowing the use of alternative standards or by specifying a range of measures, varying by grade, subject area, or, in one case, by pretest score. All four of the States using complex measures failed to meet the requirements of the law and were cited with a finding during their SPR. Three of the four States have since revised their standards to meet the Federal standards. The fourth State is in the process of revising the standard, The complex measures were illegal because their standards allowed a school with aggregate performance losses to meet the State standard while not meeting the requirements of the Federal statute.

From these data, the following observations can be made:

- more than half (62.8%) of the jurisdictions have adopted the minimum standard of aggregate performance in the Chapter 12 legislation.

- over a quarter (27.9%) of the jurisdictions have chosen to place a higher standard than that required by the Federal legislation.

- the simpler the criteria is for aggregate performance, the more likely it is to be in compliance with the legislative provisions of Chapter 1. (MacDonald, 1991, page 5).

Closer inspection of the tables in the report reveals that at least two States have used something other than the norm referenced standardized tests called for in Section 200.82 of the Rules and Regulations. One used only the advanced skills subtest and another is described as follows: "In grades where a norm-referenced test cannot be used, the performance must meet the program objective." (MacDonald, 1991, page 7).

Thus, with the possible (but unlikely) exception of this one last State, none of the 43 jurisdictions reviewed is attempting to use a performance assessment in its evaluation of Chapter 1.

H. A time for rethinking Chapter 1 requirements.

The reason this is a particularly good time to raise concerns with the Federal policy placing exclusive reliance on norm-referenced standardized tests is that there is new Federal law requiring the U.S. Department of Education to establish a committee to look at this policy and the tests used to assess achievement in Chapter 1.

Last year, the 101st Congress passed H.R. 3910, know as the "1992 National Assessment of Chapter 1 Act". The Act requires the Secretary of Education to establish an independent review panel and conduct a thorough study of Chapter 1. Among the particulars required for the study is a provision calling for --

> Descriptions and evaluations of...the overall operation and effectiveness of part A of chapter 1, including... program administration, particularly... the adequacy of standardized tests. (Section 2(b)(3)(C) of H.R. 3910, P.L. 101-305, 104 Stat. 253).

Moreover, there is another reason to hope that this is a good time to examine alternatives. The new Secretary of Education brings considerable experience with testing issues with him as he provides leadership to the Department. Discussing Federal policy in testing and assessment in the area of the National Assessment, the Study Group, chaired by then Governor of Tennessee and President of the National Governors Association, Lamar Alexander, had this to say:

> [T]he development of skilled and flexible thinking does not need to wait upon the mastery of more "basic" or "fundamental" skills grounded in rote memorization. Recent evidence indicates that young children are able to bring some of these higher processes of thought to problem solving when the tasks do not place too heavy demands upon their more limited memory skills. Research findings also show that not all subject areas require the same types of thinking skills.
>
> The national assessment should also use new technologies to develop assessment methods that go beyond the limitations of the standard multiple-choice format. Multiple-choice examinations may be easier to score and more economical to administer. But they do not easily highlight and measure those higher-order skills that lead up to and organize simpler skills...(Alexander and James, page 16).

Moreover, this is not a one time involvement of the new Secretary in this issue. From the meetings, materials and discussion seen thus far, it is clear already that we can expect some additional encouraging language for improved assessment procedures from the Mathematics and Sciences Education Board committee that Lamar Alexander has chaired. The Education Secretary also became familiar with these issues during his tenure as chair of the Assessment Planning Committee, which was the governing body for the National Assessment prior to the creation of the National Assessment Governing Board in 1986.

I. Criticism of the current evaluation policy.

What, then, are the problems with the current policy for Chapter 1 program evaluation? While there are quite a few very serious problems, each of which requires attention, let me limit the list to four points:

- Relying exclusively on <u>any single</u> outcome measure to evaluate local Chapter 1 programs is a bad idea. What have we learned from evaluation? (Scriven, 1976) ...or from education indicators for that matter? (Oakes, 1986; Murnane & Raizen, 1988)

- Norm-referenced standardized tests, as a group, do not deserve a Federally imposed monopoly. Most importantly, there are other tests that are just as good, especially in respects considered important at the local level. Other tests could satisfy Federal standards that are at least as academically stringent as the current ones (i.e., Participants must do as well as comparable students not receiving Chapter 1 services). Furthermore, not all norm referenced standardized tests are the same -- in the quality of their question, in the quality of their norms or in the utility of the information they provide. There is no justification for carte blanche approval and requirement of this class of tests.

- Nationally representative aggregate measures of program effectiveness for Chapter 1 needed for Federal level program administration and reauthorization can be obtained with less burden by using small periodic national studies rather than aggregating up result from all five million plus students participating in local Chapter 1 programs.

- The dominance of the multiple choice format in testing has produced negative impacts in many areas of education, including: curriculum and instruction, the role of teachers, and our very understanding of what students know and can do. The reliance the education community has placed on these tests has come at a very high cost.

J. Why should some performance assessments be available for

use in Chapter 1 programs?

It has been reported elsewhere (National Commission on Testing and Public Policy, 1990) that students in the U.S.A. are tested more than in any other nation. It has also been observed that all the testing does not seem to be doing much to bolster our achievement.

Moreover, Chapter 1 students as a group must be if not the most tested group certainly one of the most tested. The sheer amount of testing these students experience should give some policy makers pause for thought: These students are given standardized multiple-choice tests to define their eligibility for getting into the Chapter 1 program and again for the pretest and posttest measures currently required for annual local Chapter 1 program evaluations. Such unaltered repetition of the same types of questions in the same format is bound to give a direct message to the student that providing answers to multiple-choice items is what it is all about. If they don't get the message from the test experiences directly, don't worry , there will be plenty of other opportunities. Chapter 1 testing for program evaluation is "high stakes"; there are real consequences for the classroom teachers, aides, instructional specialists and counselors, principals, and district Chapter 1 program coordinator. Is there anyone here who doubts that these actors know the importance of improved performance on the tests and employ instructional that stress getting additional items correct?

Beyond the repetition and the high stakes pressure to do well on these test, there are problems in the interaction between these test forms and the Chapter 1 students who take them (Wolf, 1990). Chapter 1 students are not a representative sample of students in America. They are a subpopulation or more accurately a collection of subpopulation with special characteristics. Some of these problems are: test norms are less valid and less robust; there is a continuing debate even about what forms to give Chapter 1 students and which norms to apply; the limited validity of multiple choice tests is likely even more pronounced for special subpopulations; low achieving students are not likely to become engaged in repetitive, uninteresting items; and most importantly -- cognitive, language and cultural factors can provide alternative explanations probably closer to the truth than standard interpretations.

Because these kids need additional help, it is even more important that teachers be able to use the results of assessments, but the unfortunate irony is that the results are more difficult to use. There is also evidence that Chapter 1 teachers and classroom aides are generally less rather than more experienced. They need additional training in setting standards and

interpreting performance.

Chapter 1 students need additional help. Performance assessments provide some of the needed help through greater connection between curriculum, assessment activities and classroom instruction. Performance assessment items are inherently more interesting, more engaging, more thought provoking and more varied, provide better feedback on how to improve performance.

High achieving students may spontaneously integrate pieces of information or see how concepts can be applied in different situations. Low achieving students are less likely to do so. They need curriculum and instruction that helps to make those connections and applications, and they are not going to get it if the incentives and stakes are against it.

K. In search of a better array, balance and use of item formats

What is the right balance between various item types. This is not a technical question and should not be answered on technical grounds. Part of the answer should come from subject matter considerations. Some from teaching perspective. And some from resource considerations.

Current day test construction and scaling are designed to produce an aggregate score, possibly with subscales. Individual items are seen as a means to that one single end, despite the elaborate process of setting forth an array of assessment objectives. Items are initially constructed, and later selected, rejected, modified and pruned on the basis of their contribution to reliability, coherence around a single concept, and capacity to discriminate between high and low achieving students in the same way that the other items in the test do. It was in this context that Bob Mislevy observed that "We are using 20th Century mathematics to measure 19th Century psychological concepts."

What is the right balance between item types? There are several examples around that could be examined with regard to the quality of measurement, resource requirements, and the best impact on the instructional process itself and on the key actors in the educational system such as students and teachers.

Consider writing assessment (Spandel & Stiggins, 1980). About a decade ago there was a major swing in the way that writing is assessed in large scale assessments. Up to that time, writing was assessed almost entirely through multiple choice examinations in the same manner other subjects

are still assessed most places today. Yet there are still major problems in developing reliable measures of writing proficiency, especially at the individual student level. Writing performance seems very much to depend on which writing prompts are given and on anchoring procedures needed to assess trends. Also, individual students do better or less well, depending on the particular writing prompt, so that rank order is a problem across different prompts. Why is it that writing can be assessed entirely by direct assessment while other subjects are assessed entirely by multiple choice?

It is often said that multiple choice tests cost less than performance assessments. Probably so. They take less student and teacher time to achieve a reliable measure of achievement. The comparison is not that simple, however. How much is enhanced validity and impact worth?

How should assessment related issues be handled in a fair cost comparison? It is very likely that multiple choice tests mask the need for a new and more explicit, more integrated form of standard setting in classrooms. Also, Tomlinson (1988) reported that class size made no appreciable difference on student achievement. Would the same conclusion hold with performance assessment measures rather than norm referenced standardized tests? Is Harold Stevenson (1991) right in observing that we may need to give our teachers preparation time comparable to teachers in Pacific Rim countries if we aim to have comparable instruction and student achievement?

A decrease in class size or an increase in teacher preparation time may be needed to make higher standards and performance assessments operational. Either would likely spell higher costs. Should such costs be part of a comparison of multiple choice and performance assessments?

These are the types of questions that eventually we will need to deal with. In the context of Chapter 1 evaluations, however, the issues of achieving an appropriate balance are less complex. What does the Federal government need for monitoring Chapter 1 and what does it need for measuring program effectiveness? And what is a reasonable amount of discretion to leave to States and LEAs for administering and evaluating local programs?

Bibliography

Alexander, L. & T. James (1987). The Nation's Report Card. Cambridge, MA: National Academy of Education.

Chapter 1 Policy Manual: Basic Programs Operated by Local Educational Agencies. (1990). U. S. Department of Education. Washington D.C.

Bransford, J. D. & N. J. Vye (1989). "A Perspective on Cognitive Research and its Implications for Instruction" in Toward the Thinking Curriculum: Current Cognitive Research. (1989 Yearbook of the Association for Supervision and Curriculum Development). Arlington, VA.

Fredericksen, J. R. & A. Collins (1989). "A Systems Approach to Educational Testing." Educational Researcher, 18(9): 27-32.

Madaus, G.(1988). "The Influence of Testing on Curriculum" in Critical Issues in Curriculum, 87th Yearbook of the National Society for the Study of Education, Part I. Chicago, IL: Chicago Press.

MacDonald, J. T. (1991). (Memorandum) Draft Annual Report on Program Improvement [for ESEA Chapter 1]. U. S. Department of Education. Washington, D.C.

Murnane, R. J. & S. A. Raizen (eds.) (1988). Improving Indicators of the Quality of Science and Mathematics Education in Grades K - 12. Washington D. C.: National Academy Press.

National Commission on Testing and Public Policy (1990). From Gatekeeper to Gateway: Transforming Testing in America. A Report of the National Commission on Testing and Public Policy.

National Council of Teachers of Mathematics (1987). Curriculum and Evaluation Standards for School Mathematics. Reston, VA: National Council of Teachers of Mathematics.

Oakes, J. (1986). Education Indicators: A Guide for Policy Makers. Santa Monica, CA: The RAND Corporation.

Pelavin Associates (1990). Performance Assessments in the States. Washington D.C.

Resnick, L. B. & L. E. Klopfer (eds.) (1989). Toward the Thinking Curriculum: Current Cognitive Research. (1989 Yearbook of the